UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

December 28, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #V-1

MEMORANDUM FOR Dennis Stoudt
Assistant Division Chief, Processing and Support
Decennial Systems and Contracts Management Office

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Michael Starsinic/Jae Kim MS JK
Long Form and Variance Estimation Staff

Subject: Computer Specifications for Variance Estimation for Census 2000

The attachment contains the computer programming specifications for calculating the total variance at all geographic levels. These specifications are a first version of the final specifications that will be prepared on this subject. The specifications should be used to flowchart the process, to generate discussion on the requirements, to identify and finalize the record layouts of input and output files, to write computer software to implement and test the methodology. During and after the testing phase, it is likely that specifications changes will be necessary to reflect final data processing plans and other anomalies or unanticipated issues. Please direct any questions about these specifications to Michael Starsinic at (301) 457-1638.

These specifications assume:

- All correct enumeration and match probabilities have been assigned,
- All phases of the P and E-Sample missing data process have been completed, and
- Poststrata DSE estimates have been computed.

Additional specifications will be provided for producing generalized variance estimates.

Attachment

cc:

DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List

COMPUTER SPECIFICATIONS FOR VARIANCE ESTIMATION FOR CENSUS 2000

I. INTRODUCTION

A. Variance Components

Replication methods will be used to estimate the variance due to Accuracy and Coverage Evaluation (A.C.E.) sampling and estimation. These variance estimates will reflect three components of the variation of the census estimates:

1. Variance due to the multi-stage sampling of block clusters for the A.C.E., excluding the large block subsampling operation.

2. Variance due to sampling for the Targeted Extended Search (TES).

3. Variance from estimating the missing data in A.C.E.

B. Contents

This specification describes the variance estimation process for the Census 2000 Dress Rehearsal. It contains two major parts. These are:

- Overview of A.C.E. Sampling and Estimation, and TES
- Computation of the A.C.E. Variance

II. OVERVIEW OF A.C.E. ESTIMATION, A.C.E. SAMPLING and TES

A. A.C.E. Sampling

Block clusters were initially stratified by size of cluster and whether the cluster was on an American Indian Reservation (AIR): 0-2 housing units ("small" block clusters), 3-79 housing units and not on an AIR ("medium" block clusters), 80 or more housing units and not on an AIR ("large" block clusters), and 3 or more housing units on an AIR ("AIR" block clusters; not all states have this stratum). Systematic samples were selected independently in each state.[1]

Medium and large block clusters were subsampled to lower the total number of housing units in the A.C.E. sample to approximately 300,000. They were stratified based on minority/non-minority status crossed by three consistency codes (although these six strata were collapsed into fewer in some states). AIR block clusters and Puerto Rico block

[1] DSSD Census 2000 Procedures and Operations Memorandum Series R-3, "Accuracy and Coverage Evaluation (A.C.E.) Survey: Block Cluster Sample Selection Specification

clusters were not subsampled in this operation.[2]

Finally, small blocks were also subsampled in a separate operation from the A.C.E. reduction sample above.[3]

B. A.C.E. ESTIMATION

Estimation of the total population of the United States is achieved by summing dual system estimates (DSE's) of mutually exclusive and exhaustive poststrata. Definitions of the poststrata for Census 2000 are forthcoming, but the DSE equation remains the same. For each collapsed poststratum, i':[4]

$$CFC_{i'} = \frac{C_{i'} - IIC_{i'}}{C_{i'}} \times \frac{CE_{i'}}{E_{i'}} \times \frac{P_{N,i'} + P_{I,i'} \times ADJR_{2,i'}}{M_{1,i'} + \frac{M_{2,i'}}{P_{O,i'}} \times P_{I,i'} \times ADJR_{2,i'}}$$

$$DSE_{i'} = CFC_{i'} \times C_{i'}$$

where : $CFC_{i'}$ = coverage factor
$C_{i'}$ = unweighted census count
$IIC_{i'}$ = count of not data defined and wholly imputed persons
$CE_{i'}$ = estimated number of A.C.E. E-Sample persons
$E_{i'}$ = estimated number of A.C.E. E-Sample correct enumerations
$P_{N,i'}$ = estimated number of A.C.E. P-Sample nonmovers
$P_{I,i'}$ = estimated number of A.C.E. P-Sample inmovers
$P_{O,i'}$ = estimated number of A.C.E. P-Sample outmovers
$M_{1,i'}$ = estimated number of A.C.E. P-Sample nonmover matches
$M_{2,i'}$ = estimated number of A.C.E. P-Sample outmover matches
$ADJR_{2,i'}$ = ratio of estimated number of A.C.E. P-Sample outmovers to sum of P Sample noninterview adjusted weights

Puerto Rico, while handled separately, follows the same estimation process.

Missing data must be imputed for items needed for the estimation process. Item nonresponse is imputed through either hot deck or ratio methods. Unresolved residence and match status in the P-sample, and unresolved correct enumeration status for the E-

[2] DSSD Census 2000 Procedures and Operations Memorandum Series R-*, "Accuracy and Coverage Evaluation Survey: Reduction Specification"

[3] DSSD Census 2000 Procedures and Operations Memorandum Series R-*, "Accuracy and Coverage Evaluation Survey: Small Block Cluster Subsampling"

[4] DSSD Census 2000 Procedures and Operations Memorandum Series Q-*, "Dual System Estimation Computer Specifications for Census 2000"

sample are also imputed using ratio methods. Households selected but not interviewed in the A.C.E. are compensated for by using a noninterview weight adjustment.[5]

C. TES SAMPLING

For the A.C.E., 20 percent of clusters will have their surrounding blocks searched for additional matches and correct enumerations of persons who may have been misplaced geographically, as opposed to 100 percent of clusters in the 1990 Post-Enumeration Survey. Approximately 10 percent will be selected with certainty, including relist clusters and clusters with high numbers of weighted or unweighted nonmatches and geocoding errors. The remaining 10 percent will be taken from a systematic sample of the remaining clusters that contain at least one "interesting housing unit" (a unit coded in Initial Housing Unit Matching as a geocoding error, or a housing unit on the Independent List which did not match to a census unit).[6]

III. COMPUTATION OF THE A.C.E. VARIANCES

A. Overview

For Census 2000, a standard stratified jackknife such as was used in the Census 2000 Dress Rehearsal cannot be used because it cannot properly take into account the variance due to TES. A modification of the Rao-Shao jackknife variance estimator for a reweighted expansion estimator developed by Jae Kim will be used instead.[7]

The E and P-Sample A.C.E. person records are ordered in sampling strata x block cluster order. In the jackknife, a block cluster from stratum k with n_k sample block clusters is deleted from one replicate, while the estimates from the remaining block clusters in the stratum are multiplied by $n_k/(n_k-1)$.

For each replicate, the following steps of estimation should be performed and implemented:

1. Recalculation of the missing data adjustments for missing P-sample match status, E-sample correctness of enumeration, and residence probability

[5] Ikeda, Michael. Internal memorandum, "Overview of Proposed Missing Data Procedures for the Census 2000 Accuracy and Coverage Evaluation Sample"

[6] DSSD Census 2000 Procedures and Operations Memorandum Series R-20, "Accuracy and Coverage Evaluation Survey- Identification and Sampling of Block Clusters for Targeted Extended Search"

[7] Kim, Jae Kwang. Internal memorandum, "Replication Method for Two-Phase Sampling"

2. Calculation of the dual-system estimate

After the jackknifing is finished, the final step is to create a variance-covariance matrix based on the collapsed poststrata. Because of this last step and the jackknife processing, it is recommended that the variance estimation described in this specification be carried out in VPLX.

B. Input Files

1. Collapsed Poststratum File

This file contains the poststrata collapsing patterns and population and IIC counts for each.

2. Sample Design File

This file is needed to assign collapsed sampling strata to clusters.

3. Missing Data Files

Output files produced in P & E-Sample Missing Data Processing.

C. Output Files

1. A.C.E. and TES Replicate Weight Files

This file contains the replicate weights for both the A.C.E. and TES jackknifes.

2. A.C.E. Geographic Variance File

This file contains 53 variances - one national total, one for each state, and one each for the District of Columbia and Puerto Rico.

3. A.C.E. Poststrata Variance File

This file contains the national-level variance for each collapsed poststratum.

4. A.C.E. Variance-Covariance Matrix

This file contains an I x I variance-covariance matrix (where I is the number of collapsed poststrata). This file will be crucial in computing generalized variances, which will be described in a forthcoming specification.

D. Creating the A.C.E. Replicate File

1. Create the Collapsed Poststrata File

Following the completion of A.C.E. estimation, create a file showing how the poststrata were collapsed for estimation. Also include the initial phase and IIC (non-data-defined persons and persons from whole household imputations) estimates for each poststratum. The example below shows that poststratum 06111 was collapsed into 06121, and 06112 was collapsed into 06122.

	COLLAPSED		
POST-	POST-		
STRATUM	STRATUM	$C_{IP,i}$	$IIC_{IP,i}$
06111	06121	1245	123
06112	06122	23	0
06121	06121	765	34
06122	06122	564	10
06131	06131	23986	1111
etc.			

2. Create the Final Sampling Stratum Variable

From the Sample Design File, extract the block cluster records of all clusters in the A.C.E. sample after A.C.E. Reduction and Small Block Subsampling.

For each block cluster record, extract the following variables:

VARNAME	Variable Description	Location in File
STATE	FIPS State Code	3-4
CLUST	A.C.E. Block Cluster Number	21-25
DIGIT	A.C.E. Block Cluster Check Digit	26
SS	Sampling Stratum: 1 = Small 2 = Medium 3 = Large 4 = Medium or Large on AIR	55
ARS	A.C.E. Reduction Stratum	190-191
SBCSS	Small Block Cluster Sampling Stratum	306-307

Create a new variable, Final Sampling Stratum (FSS):

$$FSS = SBCSS + 100*ARS + 10000*SS + 100000*STATE$$

3. Collapse Final Sampling Strata

It is possible that an FSS may contain only one block cluster. In this situation the sampling stratum needs to be collapsed into another stratum.

a. For each final sampling stratum, count the number of block clusters.

b. If any final sampling stratum contains one block cluster, recode the value as follows:

[Collapsing algorithm to be determined.]

c. After all affected clusters have had their stratum recoded, recompute the number of block clusters in each sampling stratum.

Define: S = number of collapsed sampling strata in the nation
k = sampling stratum index; $1, \ldots, S$
n_k = number of block clusters in sampling stratum k

4. Create the Cluster-Level A.C.E. and TES Replicate Weight File

Replicate weights are the means by which the jackknife is implemented. Two separate sets of replicate weights will be needed, as the variance formula requires separate jackknifes based on the A.C.E. weights (RWA) and the TES weights (RWT).

For the A.C.E. replicates, there are N+1 sets of replicate weights, one for each of the N clusters and one for the full sample. For the 0th (full sample) replicate, all replicate weights are equal to one. For the jth replicate, the replicate weight for the jth cluster is equal to zero, the replicate weights for all other clusters in that sampling stratum is $n_k / (n_k\text{-}1)$ where n_k is the number of clusters in that sampling stratum, and the replicate weights for all clusters not in that sampling stratum are equal to one.

For the TES replicates, there are N_t+1 sets of replicate weights, one each for the N_t clusters selected in the systematic portion (not the certainty portion) of the TES sample and one for the full sample. For clusters ineligible for selection in the TES sample and clusters selected in the TES sample with certainty, set all their replicate weights to 9 (these will not be used in computation, and are meant to differentiate them from eligible clusters with a replicate weight of zero). For the remaining clusters (those selected in the systematic sample for TES), for the jth replicate, the replicate weight for the jth cluster is equal to zero, and the replicate weights for all other eligible clusters is $N_t / (N_t\text{-}1)$.

All persons in a cluster receive the same weight.

5. Create the E- and P-Sample Replicate Files

a. Extract the following variables from the E-Sample Person Missing Data Output File:

[Variable list forthcoming. Whatever variables are needed to determine poststrata membership, plus necessary weights, sampling stratum identifiers, and geographic information.]

Extract the following variables from the P-Sample Person Missing Data Output File:

[Variable list forthcoming. Whatever variables are needed to determine poststrata membership, plus necessary weights, sampling stratum identifiers, and geographic information.]

b. Append collapsed FSS from step 3 (merge by A.C.E. Cluster Number and Check Digit) to both files.

c. Assign individuals in both files to collapsed poststrata using the same algorithm that was used in A.C.E. estimation. [Algorithm forthcoming.]

d. Append both sets of replicate weights (RWA and RWT, merged by cluster) to both the E- and P-Sample files.

E. Compute the A.C.E. Variances

1. Replicate Missing Data Imputation

Adjust CEPROBF on the E-Sample File, and MPROB and RPROB on the P-Sample file based on the replicate weights. [A much more explicit description of this step is forthcoming.]

2. Replicate DSE Estimation and Compute Variances

a. On the E-Sample file, sum the following values for each FSS x CLUST x CPS, separately for non-TES persons, TES persons in certainty blocks, and TES persons in non-certainty blocks (T^* is equal to TESWGT if the person is a TES person and is equal 1 to otherwise):

EWGHT x T^*
CEPROBF x EWGHT x T^*

On the P-Sample file, sum the following values for each FSS x CLUST x CPS, separately for non-TES persons, TES persons in certainty blocks, and TES persons in non-certainty blocks:

RPROB x NIWGTO x T^* for nonmovers only
RPROB x NIWGTO x T^* for outmovers only
RPROB x NIWGTI x T^* for inmovers only
MPROB x RPROB x NIWGTO x T^* for nonmovers only
MPROB x RPROB x NIWGTO x T^* for outmovers only
NIWGTO x T^* for outmovers only

b. The basic formula for DSE estimation is:

$$DSE_{i'} = (C_{i'} - IIC_{i'}) \times \frac{CE_{i'}}{E_{i'}} \times \frac{P_{N,i'} + P_{I,i'} \times ADJR_{2,i'}}{M_{1,i'} + \frac{M_{2,i'}}{P_{O,i'}} \times P_{I,i'} \times ADJR_{2,i'}}$$

where:

$$E_{i'} = \sum_{j \in i'} EWGHT_j \times T_j^*$$

$$CE_{i'} = \sum_{j \in i'} CEPROBF_j \times EWGHT_j \times T_j^*$$

$$P_{N,i'} = \sum_{j \in i', j \in nonmover} RPROB_j \times NIWGTO_j \times T_j^*$$

$$P_{O,i'} = \sum_{j \in i', j \in outmover} RPROB_j \times NIWGTO_j \times T_j^*$$

$$P_{I,i'} = \sum_{j \in i', j \in inmover} RPROB_j \times NIWGTI_j \times T_j^*$$

$$M_{1,i'} = \sum_{j \in i', j \in nonmover} MPROB_j \times RPROB_j \times NIWGTO_j \times T_j^*$$

$$M_{2,i'} = \sum_{j \in i', j \in outmover} MPROB_j \times RPROB_j \times NIWGTO_j \times T_j^*$$

$$ADJR_{2,i'} = \frac{\sum_{j \in i', j \in outmover} RPROB_j \times NIWGTO_j \times T_j^*}{\sum_{j \in i', j \in outmover} NIWGTO_j \times T_j^*}$$

If the denominator of any ratio is equal to zero, set that ratio equal to 1.

c. There are three components to the variance which must be estimated separately. Each is based on the slightly different application of the above formula.

Component 1

$$\hat{V}_r = \sum_{k=1}^{S} \frac{(n_k - 1)}{n_k} (D\hat{S}E_r^{(k)} - D\hat{S}E_r)^2$$

$$\hat{Term}_r^{(k)} = \sum_{i \in A} \sum_{j} w_{ij}^{(k)} m_{ij} x_{ij} + \sum_{i \in A} \sum_{j} w_{ij}^{(k)} m_{ij} y_{ij}$$

$$+ \frac{\sum_{i \in A} \sum_{j} w_{ij}^{(k)} z_{ij}}{\sum_{i \in A} \sum_{j} w_{ij}^{(k)} t_{ij} z_{ij}} \sum_{i \in A} \sum_{j} w_{ij}^{(k)} t_{ij} m_{ij} z_{ij}$$

Where a "Term" is any of the eight individual numerators and denominators making up the second and third terms of the DSE formula, and:

i = cluster designator
j = person designator
A = the realized A.C.E. sample
x_{ij} = 1 if the person is NOT a TES person, 0 otherwise
y_{ij} = 1 if the person IS a TES person AND is from a cluster sampled with certainty in TES, 0 otherwise
z_{ij} = 1 if the person IS a TES person AND is from a non-certainty cluster sampled in TES, 0 otherwise
$w_{ij}^{(k)}$ = jackknife-adjusted "weight", e.g. RWA x EWGHT for E and CE, RWA x RPROB x NIWGTO/I for P's and M's
m_{ij} = "match weight", e.g. CEPROBF for CE, MPROB for M's, 1 otherwise
t_{ij} = TES weight

The first term sums the weights for non-TES persons in all clusters. The second term sums the weights for TES persons in clusters selected in TES with certainty. The third term sums the weights for TES persons in all other clusters (not all clusters will contain TES persons). For each jackknife replicate, compute each component separately and combine them to form the $DSE_r^{(k)}$. C_i and IIC_i are obtained from the collapsed poststrata file.

Component 2

$$\hat{V}_{2r} = \sum_{s=1}^{M} \frac{(N_t - 1)}{N_t} (D\hat{S}E_r^{(s)} - D\hat{S}E_r)^2$$

$$\hat{Term}_r^{(s)} = \sum_{i \in A} \sum_{j} w_{ij} m_{ij} x_{ij} + \sum_{i \in A} \sum_{j} w_{ij} m_{ij} y_{ij}$$

$$+ \frac{\sum_{i \in A} \sum_{j} w_{ij} z_{ij}}{\sum_{i \in A} \sum_{j} w_{ij} t_{ij}^{(s)} z_{ij}} \sum_{i \in A} \sum_{j} w_{ij} t_{ij}^{(s)} m_{ij} z_{ij}$$

where:

N_t = number of clusters sampled systematically (i.e. sampled but not with certainty)

w_{ij} = "weight", e.g. EWGHT for E and CE, RPROB x NIWGTO/I for P's and M's

$t_{ij}^{(s)}$ = jackknife-adjusted TES weight, i.e. RWT x TESWGT

Component 3

$$\hat{V}_{2d} = \sum_{s=1}^{M} \frac{(N_t - 1)}{N_t} (D\hat{S}E_d^{(s)} - D\hat{S}E_d)^2$$

$$\widehat{Term}_r^{(s)} = \sum_{i \in A} \sum_j w_{ij}\, m_{ij}\, x_{ij} + \sum_{i \in A} \sum_j w_{ij}\, m_{ij}\, y_{ij} + \sum_{i \in A} \sum_j w_{ij}\, t_{ij}^{(s)}\, m_{ij}\, z_{ij}$$

The total variance is estimated by:

$$\hat{V}_d = \hat{V}_r - \hat{V}_{2r} + \hat{V}_{2d}$$

d. Output the A.C.E. Geographic Variance File, A.C.E. Poststratum Variance File and the A.C.E. Variance-Covariance Matrix.